Towards Measuring the Representation of Subjective Global Opinions in Language Models

Year: 2023 | Citations: 326 | Authors: Esin Durmus, Karina Nyugen, Thomas Liao, Nicholas Schiefer, Amanda Askell

Abstract

Large language models (LLMs) may not equitably represent diverse global perspectives on societal issues. In this paper, we develop a quantitative framework to evaluate whose opinions model-generated responses are more similar to. We first build a dataset, GlobalOpinionQA, comprised of questions and answers from cross-national surveys designed to capture diverse opinions on global issues across different countries. Next, we define a metric that quantifies the similarity between LLM-generated survey responses and human responses, conditioned on country. With our framework, we run three experiments on an LLM trained to be helpful, honest, and harmless with Constitutional AI. By default, LLM responses tend to be more similar to the opinions of certain populations, such as those from the USA, and some European and South American countries, highlighting the potential for biases. When we prompt the model to consider a particular country's perspective, responses shift to be more similar to the opinions of the prompted populations, but can reflect harmful cultural stereotypes. When we translate GlobalOpinionQA questions to a target language, the model's responses do not necessarily become the most similar to the opinions of speakers of those languages. We release our dataset for others to use and build on. Our data is at https://huggingface.co/datasets/Anthropic/llm_global_opinions. We also provide an interactive visualization at https://llmglobalvalues.anthropic.com.